图卷积网络谣言检测

xxx

xxx

摘 要: 随着社交媒体的普及, 网络谣言的传播已经成为一个严重的社会问题, 本文旨在提出一种高效的网络谣言检测方法, 以应对社交媒体上不实言论的传播。通过分析大规模社交媒体数据集, 构建了一个谣言知识图谱, 结合图神经网络和文本嵌入技术, 实现对网络谣言的自动识别。该模型考虑了文本内容特征和用户关系信息, 更全面地捕捉了谣言传播的模式。实验证明, 该方法在谣言识别方面取得了显著的性能提升, 并具有较强的泛化能力。这一研究对于维护网络信息的真实性和社交媒体的健康发展具有重要意义。

关键词: 网络谣言检测; 主题分析; 情感分析; 图卷积神经网络;

1 引 言

近年来, 随着社交媒体的普及, 网络谣言的传播已经成为一个严重的社会问题。谣言, 即 “信息流传中的尚未得到证实的陈述”[1] 。网络谣言, 即 “通过网络介质传播的谣言”。网络谣言传播速度快, 范围广, 加之 “语不惊人死不休”, 极易蛊惑人心, 加剧社会恐慌, 破坏社会正常秩序。就比如, “支付宝、微信支付个人收款码将于明年 3 月 1 日起被禁止商用”, 事实上, 央行新规明确强调了个人静态收款码原则上禁止用于远程非面对面收款。而非所传的禁用支付宝和微信。所以在高信息量, 高流通速度的网络时代, 准确识别谣言并处罚谣言传播者将会格外重要。

针对上述问题, 本文提出了一种高效的图卷积神经网络谣言识别方法, 以应对信息传播中的不实言论。我们通过分析社交媒体上的大规模数据集, 利用了评论/转发之间的关系信息, 构建了网络谣言传播图, 结合图神经网络和文本嵌入方法, 从而更全面地捕捉网络谣言传播的模式, 实现对网络谣言的自动识别。这一方法对于维护网络信息的真实性和社交媒体的健康发展具有重要意义。

本文以微博平台为例, 对网络谣言展开分析, 主要贡献有三方面: 第一, 本文对微博网络谣言进行了情感分析和主题分析, 对谣言的情感极性和主要的主题做出了归类。第二, 本文借助文本传播关系, 建立了网络谣言传播的知识图谱, 分析了谣言和非谣言在传播结构上的区别。第三, 本文利用所构建的知识图谱, 使用图卷积神经网络, 实现了对网络谣言的识别, 且有较强的泛化能力, 可以节省用于网络谣言识别的人力物力, 对于改善网络环境具有重要意义。

2 相关研究

社交网络的迅速发展使得以微博、微信等为代表的线上社交媒体逐渐取代传统媒体, 人们面对海量的信息, 难以准确判别信息的真伪, 加之社交网络平台对信息发布的监管不力, 更易造成谣言泛滥。故谣言的本质属性在于难以确定信息的真实性, 本文将网络谣言定义为在社交媒体平台上广泛传播的、信息真实性未确定的信息陈述。针对于社交网络中谣言泛滥的问题, 已经有众多学者和研究旨在寻找有效的检测谣言的方法, 以帮助人们辨别信息真伪。

现有的社会网络谣言检测主要分为三类: 人工检测方法、基于传统机器学习的检测方法与基于深度学习的检测方法。人工检测虽然准确率较高, 但是无法及时迅速检测网络中的海量数据; 传统的机器学习关注对文本特征和时序特征的挖掘, 一般根据谣言的内容、用户属性、传播方式人工地构造特征, 将问题转

作者简介:xxx

化为基于人工提取特征的分类问题，自动化程度明显提高。Ma,Gao 等 (2015)[2] 充分利用社交平台中的文本特征会随着时间推移而变化的特性，指出基于时间序列的建模方法对于谣言检测至关重要。此外,Liu 等 (2015)[3] 面向 Twitter 研究谣言检测算法，其主要思想是挑选出包含常识性知识和调查性新闻的评论，并假设它们是对相关信息真实性的争论，以此判断相关信息是否为谣言。

近年来，深度学习被广泛应用于谣言识别中，比传统的机器学习中通过特征工程得到的特征数据对原数据有更好的表征性，从而实现更好的分类效果。刘政等[4] 利用基于卷积神经网络（CNN）的谣言检测模型，将微博中的谣言事件向量化，通过卷积神经网络隐含层的学习训练来挖掘表示文本深层的特征，避免了特征构建的问题，并能发现那些不容易被人发现的特征。任文静[5] 等研究了 GRU，LSTM 等深度学习模型在谣言检测上的应用，判断微博文本是否为谣言类信息。陈志毅[6] 等提出了一种基于深度神经网络，针对配文文本内容、图像以及用户属性信息的多模态网络谣言检测方法 DCNN。实验结果表明 DCNN 算法将识别准确率从 78.1% 提高到了 80.3%，验证了 DCNN 算法和其中对社会特征建立特征交互方法的可行性与有效性。

在谣言检测的问题上，现有的研究方法无法有效地表达谣言的传播结构，并且没有引入外部知识作为内容核实的手段。因此，本文提出了引入知识图谱表示的图卷积网络谣言检测方法，其中知识图谱作为额外先验知识来帮助核实内容真实性。采用预训练好的词嵌入模型和知识图谱嵌入模型获取文本表示后，融合图卷积网络的同时，能够在谣言传播的拓扑图中更好地进行特征提取以提升谣言检测的精确率。

3 实验数据

本文第一部分数据使用 GitHub 中新浪微博谣言数据，该数据为从新浪微博不实信息举报平台抓取的中文谣言数据。数据集中共包含谣言 1538 条和非谣言 1849 条，以及其评论/转发约万余条。该数据集分为微博原文与其转发/评论内容。其中所有微博原文（包含谣言与非谣言）在 original-microblog 文件夹中，剩余两个文件夹 non-rumor-repost 和 rumor-repost 分别包含非谣言原文与谣言原文的对应的转发与评论信息（该数据集中并不区分评论与转发）。该数据文件中，每条原文，评论或评论均为 json 格式的数据。本文所使用的数据内容如下表。

表 1: 所使用谣言数据内容

变量	含义	数据类型
MID	用户的消息 ID	STRING
Friends	粉丝用户数	INT
Likes	关注用户数	INT
Parents	转发或评论的原文章 MID	STRING
Kids	转发或评论此文章 MID	STRING
Text	评论文本	STRING

第二部分数据为自己手动爬取的微博日常评论，存储在 wb_py.xls 文件中，包括 5032 条数据，数据内容包括用户和博文内容，用于 4.1 情感分析，研究主要的情感倾向。

图 1: 微博不实信息举报平台示例

4 数据处理及探索性分析

4.1 情感分析

相对于积极情感，谣言似乎与消极情感更加相关，本文从 JSON 文件中提取文本信息，随后，采用了中文分词工具 jieba，去除停用词，对每个文本进行了分词与文本清洗。为了衡量文本的情感强度，借助大连理工情感强度词典，其中包含了一系列情感词及其对应的强度，利用该词典对每个文本进行了情感分数的计算，结果如图 2。从数据来看，非谣言情感强度也偏向于消极情绪，可能的原因有两点：首先，我们所选取的数据集有 1538 条谣言数据和 1849 条非谣言数据，两种类型的谣言数据数量相近，但在实际生活中，非谣言数据明显要比谣言数据多。其次该数据集是从微博不实信息举报平台抓取的谣言数据，非谣言数据为被举报。但是

经确认是非谣言数据，也就是说，在该数据集中这些非谣言数据与谣言数据有一定的相似程度，因而不能简单的用以对比。因此，我们继续爬取了微博上的日常博文和评论，结果如下。

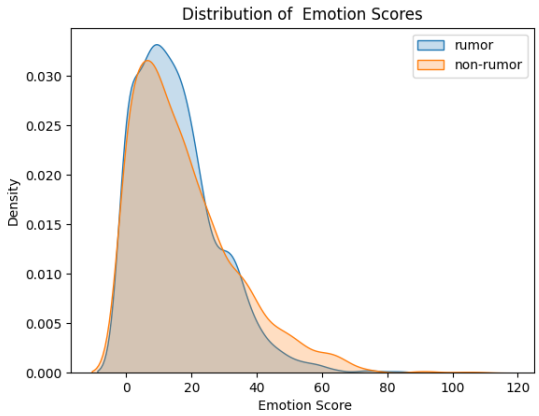

图 2: 情感强度图

因此，我们继续爬取了微博上的日常博文和评论，结果如图 3。可以看出，谣言情感确实要偏向消极。根据心理学情绪理论，消极情感能够更容易地与其他个体的负面情绪相连接，人们更容易与负面情感产生共鸣。

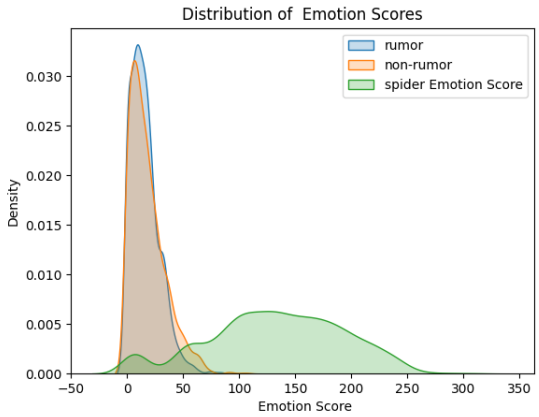

图 3: 情感强度图

4.2 主题分析

为了深入了解谣言传播网络中的主题分布，本文采用了主题建模方法。主题建模是一种从文本数据中发现主题的技术，通过挖掘文本中隐藏的语义结构，揭示文本内容的主题分布。

本研究采用了 Latent Dirichlet Allocation (LDA) 方法进行主题建模。LDA 是一种概率图模型，假设文本由多个主题组成，每个主题由一组单词表示。LDA 的基本原理是：

1. 每个文档包含多个主题，每个主题对应一定比例的词语。

2. 每个主题包含多个词语，每个词语在不同主题中出现的概率不同。

通过使用 LDA 模型，我们能够得到每个文本样本在不同主题上的分布情况，进而揭示谣言传播网络中的关键主题。

通过对谣言传播网络中的文本进行主题分析，我们得到了如下主题分布：

1. 主题 1：明星谣言

如图 6, 包括的关键词：微信、收费、女友、律师团、强奸罪、别人、酒吧等。这些关键词与明星社交活动、社会关系等相关，多半是为了蹭取明星热度获得浏览所产生的网络谣言。(由于图片较小，放大可以看清，或在文件夹 picture 中查看原图)

2. 主题 2：法律谣言

如图 7, 包括的关键词：罚、安全带、火车、处罚、交通、超速、驾驶等。这些关键词与交通安全、违规行为、法律法规等相关。

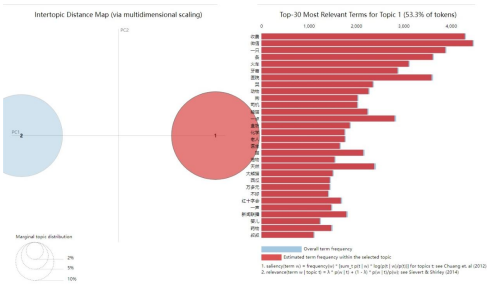

图 4: 主题 1

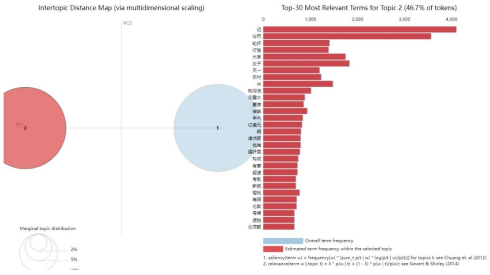

图 5: 主题 2

5 GCN 谣言检测模型的构建

5.1 文本向量化

在谣言传播网络分析中，对节点文本信息的向量化是关键的一步。本文采用了词袋模型作为文本向量化的方法。

词袋模型是一种常见的文本表示方法，它将文本看作是一个无序的词汇集合，忽略了单词的顺序，只关注每个单词的出现频率，包含以下步骤：

1. 分词（Tokenization）：已在数据探索性分析时实现

2. 构建词汇表（Vocabulary）：收集文本中所有出现过的单词，形成一个词汇表。

3. 向量化（Vectorization）：对每个文本样本，根据词汇表中每个单词的出现情况，构建一个向量表示。

5.2 网络谣言知识图谱的构建

对于每一条谣言和非谣言的数据，我们构建了对应的图谱结构。在一个信息传播图中，以原始信息为根节点，建立一个有向图，其中每个节点表示一个用户，每条有向边表示评论/转发关系。边表示了信息的传播路径，指示信息是如何从一个用户传播到另一个用户的，如图 6 ,7。每个节点具有用户属性特征，例如粉丝数量、基于信息文本内容的文本嵌入特征等，如表 2。即，对于每个图谱 G，节点集合 V，边集合 E，每个节点 $v_i \in V$，v_i 的属性为 $[MID_i, Friends_i, Likes_i, Text_i]$。

表 2: 节点属性

变量	含义	数据类型
MID	用户的消息 ID	STRING
Friends	粉丝用户数	INT
Likes	关注用户数	INT
Text	评论文本	STRING

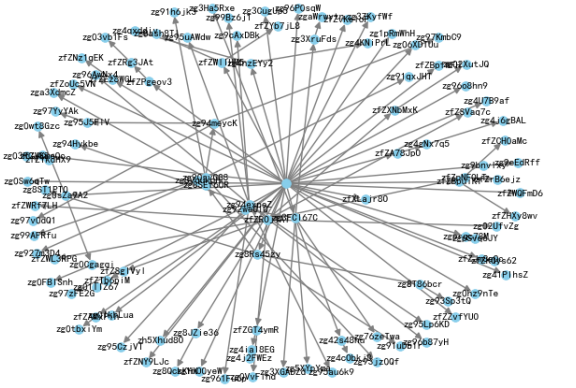

图 6: 某一条谣言的用户网络图

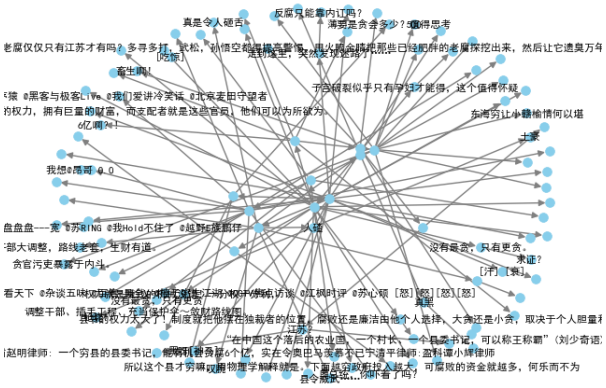

图 7: 某一条谣言的内容网络图 (图中无内容节点为仅转发不评论)

我们分别计算了谣言和非谣言图的长度（即根节点到某个叶子节点的长度），根据表 3, 谣言的平均传播长度和最大传播长度均长于非谣言。首先，谣言产生的目的往往是博取眼球，获得流量，往往包含引人注目

的内容，这可能导致更多的人转发和分享，从而使传播路径更长。其次，谣言可能引起情感反应，例如恐惧、兴奋或好奇。情感充沛的内容往往会更广泛地分享，这也导致了传播路径更长。最后，随着谣言的传播，信息往往发生失真或修改。谣言内容越传越夸张，从而吸引更多的人，导致初始源和后续节点之间的路径更长。

对于节点总数，谣言可能仅在特定社交群体内传播，就比如明星谣言的传播受众大部分为该明星的粉丝，因此，谣言图可能由于受众范围较小而具有较少的节点数量。

表 3: 传播链的平均长度

变量	传播长度	最大传播长度	节点总数
谣言	3.71	11	315.3
非谣言	2.9	9	428.8

在谣言的传播过程中，我们观察到一级节点的扩散后，二级节点数量达到峰值，这时谣言的影响力也最大。然后，随着时间的推移，谣言可能受到辟谣的影响，传播到第三级节点的数量减少。相反，对于非谣言，我们看到了一个正反馈的传播趋势，一直到热度逐渐降低，图 8。

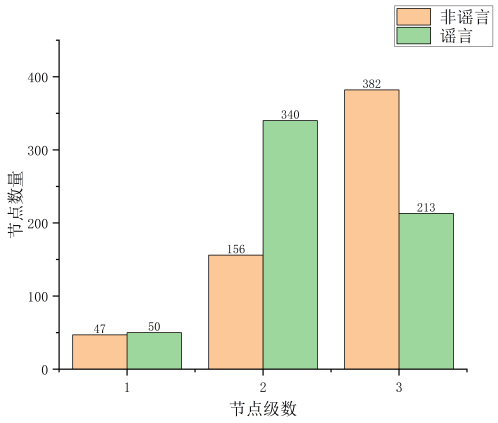

图 8: 各级平均节点数量对比

根据不同传播层级的用户属性进行的分析，表 4展示了一至三级转发/评论层次下节点的用户特征的平均情况。结果显示，随着转发层次的深入，用户的粉丝数、关注人数和博文发布数逐渐减少，尤其是粉丝数显著下降。这说明随着传播的深入，参与传播的用户更慢慢转变为网络中影响力较小的普通用户。这些用户在信息甄别和筛选方面可能相对较弱，面对不确定性和谣言事件时容易产生不同立场并引发递进式的讨论，从而使得谣言具有更深的传播结构。因此，较深的传播结构反映了微博中的不确定性和争议性。通过关注这些传播链较长的传播结构，有助于模型对谣言进行判别。

表 4: 不同转发等级特征均值

层级	平均粉丝用户数	关注用户数
一级转发/评论	32135.2	7243.2
二级转发/评论	6521.8	901.1
三级转发/评论	4785.6	853.4

5.3 GCN 谣言分类模型

5.3.1 网络架构

构建了一个基于 GCN 的模型，其中使用了两层 GCN，一个线性层，和一个 MLP 层。GCN 层用于学习节点表示，线性层用于将节点表示与输入特征连接，MLP 层用于进行最终的分类。

• GCNConv 层: GCN 层的前向传播公式可以表示为:

$$H^{(l+1)} = \sigma(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}H^{(l)}W^{(l)})$$

其中，$H^{(l)}$ 是第 l 层的节点特征矩阵，A 是邻接矩阵，D 是度矩阵，$W^{(l)}$ 是第 l 层的权重矩阵，σ 是激活函数。

• 线性层: 线性层的前向传播公式为:

$$\text{Linear}^{(l+1)} = \text{Linear}^{(l)}(H^{(l)}, H^{(0)}) = H^{(l)}W^{(l)} + H^{(0)}W^{(0)}$$

其中，$\text{Linear}^{(l)}$ 表示第 l 层的线性变换，$H^{(0)}$ 是输入的节点特征矩阵，$W^{(0)}$ 是第 l 层的权重矩阵。

•MLP 模块: MLP 模块包含三个线性层和两个 ReLU 激活函数，前向传播公式可以表示为:

$$\text{MLP}^{(l+1)} = \text{ReLU}(\text{Linear}^{(l+1)})$$

其中，$\text{Linear}^{(l+1)}$ 表示第 $l+1$ 层的线性变换，ReLU 表示修正线性单元激活函数。

•Loss 函数和优化:

交叉熵损失函数用于衡量两个概率分布之间的差异。对于分类任务，它常用于度量模型的预测分布与真实分布之间的差异。对于二分类或多分类问题，交叉熵损失函数的公式如下:

$$\text{Loss} = -\left(y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y})\right)$$

其中：y 是真实标签，$\hat{y}$ 是模型的输出

5.3.2 特征向量构建

输入为特征矩阵 H 和邻接矩阵 A，其中 H_i 表示第 i 个节点的特征向量，包含每个节点的属性 (不包括 MID)，是节点的属性信息矩阵，行数等于节点数，列数等于属性数。A 是网络结构的邻接矩阵，存储着结构信息。

6 其他对比方法

6.1 支持向量机

SVM 在处理小样本、高维度数据和复杂决策边界等方面表现出色。然而，在处理大规模数据集时，训练时间可能较长。因此使用 svm 作为评价的参考。以下是 SVM 的线性核的分类函数的一般形式：分类函数的形式为：

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right)$$

其中：$f(x)$ 是分类函数，返回样本 x 的预测类别（1 或 0）,sign(·) 是符号函数，大于等于零返回 1，小于零返回-1。α_i 是支持向量的权重系数。$K(x_i, x)$ 是核函数，对于线性核，它是 x_i 和 x 的内积。b 是偏置（截距）。

在训练阶段，通过最小化损失函数，可以得到相应的 α_i 和 b 的值。优化问题形式为：

$$\min_{\alpha,b} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{n} \alpha_i$$

约束条件为：

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C$$

其中 C 是一个正则化参数，用于平衡间隔最大化和误分类样本的惩罚。

6.2 随机森林

对于文本分类，随机森林的基本思想是通过组合多个决策树来提高整体的性能。每个决策树都是通过对不同的数据子集和特征子集进行训练得到的。能够处理非线性关系，而且相对易于解释。随机森林在处理大量的文本特征时可能表现得更加稳健[7]

6.3 朴素贝叶斯

朴素贝叶斯（Naive Bayes）是一种经典的分类算法，特别适用于处理文本信息。它的简单性和高效性使得在文本分类、垃圾邮件过滤等应用中广泛使用。

朴素贝叶斯的基本思想是基于贝叶斯定理，利用特征之间的独立性假设，计算给定类别的情况下，每个特征的条件概率。在处理文本信息时，通常用于分类任务，例如将文本分为不同的类别或标签。

7 模型评价

对模型调优之后，训练 200 代，查看损失情况，由图 9可以看出，在训练 40 代左右时，模型已到达最优。我们训练集占比 0.7，测试集预测结果混淆矩阵如表 5，大部分谣言都能被检测出来，但是对于非谣言，也有相当大一部分预测错误。

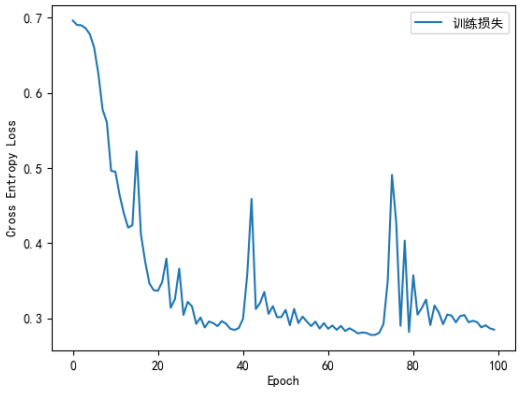

图 9: 交叉熵损失随训练代数变化情况

表 5: 混淆矩阵

	预测结果	
真实标签	谣言	非谣言
谣言	13450	203
非谣言	5183	10448

对比支持向量机、随机森林、朴素贝叶斯，结合知识图谱的 GCN 有着较好的效果。

表 6: 评价指标

方法	召回率	精确率	准确率	F1
SVM	0.881	0.693	0.752	0.746
随机森林	0.772	0.745	0.77	0.748
朴素贝叶斯	0.712	0.615	0.652	0.651
GCN	0.985	0.722	0.816	0.883

8 结论与展望

本文提出了一种基于知识图谱和图卷积神经网络的网络谣言检测方法，通过综合考虑文本内容特征和用户关系信息，取得了在谣言识别方面的显著性能提升。该模型具有较强的泛化能力，对网络信息的真实性能进行有效判断，为社交媒体上的谣言传播问题提供了一种高效的解决方案。然而，不同的转发和评论，对于谣言的传播影响是有差异的。本文未能提出一种能够客观衡量不同节点权重的方法。在图数据处理过程中，我们忽略了不同节点具有不同权重的问题，从而没有利用节点之间的多样性信息。

未来可以使用多模态信息融合，考虑融合文本信息以外的多模态信息，如图像、视频等，以更全面地理解谣言传播的背后机制。也需要进一步提高模型的实时性，使其能够应对信息传播的即时性，提高在实际应用中的实用性。

通过不断改进和拓展，网络谣言检测模型将更好地服务社交媒体信息管理，保障网络信息的真实性，促进社交媒体的健康发展。

$$\min_{h\in F} E[\ell_{01}(b,h(a))] = \min_{h\in F}\{1\cdot P(b\neq h(a))+0\cdot P(b=h(a))\} = \min_{h\in F}\{P(b\neq h($$

参考文献：

[1] DIFONZO N, BORDIA P. Rumor, gossip and urban legends[J]. Dio-genes, 2007, 54(1): 19-35.

[2] MA J, GAO W, WEI Z Y, et al. Detect rumors using time series of social context information on microblogging websites[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM Press, 2015: 1751-1754.

[3] LIU X, NOURBAKHSH A, LI Q, et al. Real-time rumor debunking on twitter[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM Press, 2015: 1867-1870.

[4] 刘政, 卫志华, 张韧弦. 基于卷积神经网络的谣言检测[J]. 计算机应用, 2017, 37(11): 3053-3056+3100.

[5] 任文静. 面向微博谣言的检测方法研究[D]. 哈尔滨工业大学, 2017.

[6] 陈志毅, 隋杰. 基于 DeepFM 和卷积神经网络的集成式多模态谣言检测方法[J/OL]. 计算机科学, 2022, 49(1): 101-107. https://doi.org/10.11896/jsjkx.201200007.

[7] 曾子明, 王婧. 基于 LDA 和随机森林的微博谣言识别研究——以 2016 年雾霾谣言为例[J]. 情报学报, 2019, 38(1): 89-96.